A Crow or a Blackbird?: Using True Social Network and Tweeting Behavior to Detect Malicious Entities in Twitter

Vijay A. Balasubramaniyan, Arjun Maheswaran, Viswanathan Mahalingam,
Mustaque Ahamad, and H. Venkateswaran

Georgia Tech Information Security Center
Georgia Institute of Technology
{vijayab, amaheswa, vmahali1, mustaq, venkat}@cc.gatech.edu

ABSTRACT

The growing popularity of Twitter and its ability to enable near instantaneous sharing of information has made it a target of attacks by malicious entities who use it to spam and provide links to malware. There is evidence that these entities are using increasingly sophisticated techniques that mimic the behavior of reputed sources to avoid detection. We use novel mechanisms that utilize the true social network of users, the quality of information produced by them and their tweeting behavior to identify such entities. A scheme based on these mechanisms is even able to detect malicious entities that collude to establish dense social networks. Using actual data from a representative sample of 278, 758 Twitter users, we demonstrate the effectiveness of this approach by showing that (1) we identified 5334 accounts that had links to unsafe websites, and (2) over a period of 31 days, 181 accounts that our algorithm identified as potentially malicious were subsequently suspended by Twitter. We believe our algorithm is one of the first to automatically deal with a broad range of malicious entities present in Twitter.

1. INTRODUCTION

Nowadays, there is a tremendous amount of information that is available from online sources and new information is constantly created and made available by these sources. Search engines help find relevant information after it has been created but the desire to have near instantaneous access to information has led to micro-blogging services such as Twitter [10]. Twitter allows users to post limited length messages or *tweets* that alert users when new information, that is of interest to them, becomes available. The popularity of entities like CNN breaking news [2] for the latest news headlines and Woot [14] for valuable shopping deals shows Twitter's effectiveness in enabling quick and easy access to fresh information. Furthermore, since Twitter started out as a social networking site, people use it to stay aware of the happenings in the lives of their friends, making it a one stop source for information about personal and public events.

The growing popularity of Twitter, with 75 million users [3], has made it an attractive target for malicious entities. Such entities inject noise that is motivated by profit or for propagating misinformation. Unlike the web, where information sources have an implicit quality based on where they appear in search engine results, in Twitter, there is no easy way of assessing if a source is legitimate or malicious. The ease with which a large number of tweets can be generated combined with no existing mechanism to establish the legitimacy of tweet sources has enabled malicious entities to exploit this medium. For example, many of these malicious entities provide links to sites that contain malware or are spammers who try to sell questionable products such as *The Greatest Vitamin in the World* [5]. Clearly, it is important to separate malicious entities from sources of useful information to counter or mitigate the denial-of-information attacks that result from noisy or misleading tweets.

Past research has explored how to characterize Twitter users into information sources, friends and information seekers [20]. In this paper, we focus on developing user characterization that helps in distinguishing malicious entities from legitimate users. This is challenging because many of the observable characteristics of malicious entities are similar to good users. The Twitter webpage for a user provides, among other things, the user's name, a short bio, a geographical location, the friend and follower count and the number of tweets. Friends are users that you follow while followers follow you. The follower and friend count can be inflated to any value through a variety of techniques [12]. In fact, many of the malicious entities that we observed had over 5000 friends and followers. The information other than follower and friend count is provided by the user and cannot be validated. Since there is a large number of such malicious entities and they can collude, there is an increased chance that someone relying on Twitter to find a health supplement ends up buying a potentially fraudulent product instead.

In this paper, we create a novel scheme that uses both the true social network structure of users and their behavioral characteristics to determine whether a certain user is legitimate or malicious. The social network (SN) of a Twitter user as advertised by the friend and follower count does not indicate real relationships [19]. Instead, we use @-messages, that direct conversation from one user to another, and retweets, that are used to disseminate information, to determine a user's SN and call it his *true social network*

(TSN). We use PageRank [25] with the TSN to compute a reputation value for each user which represents the user's ability to either engage other users or provide information that is of interest to others. As malicious entities have low quality information, a large number of them will end up with low reputation values.

A serious problem with using only PageRank, even when it is used with the TSN, is that along with malicious entities, users with moderate to small sized friend circle who use Twitter primarily as a social networking site will have low reputation values. These users tweet occasionally and most of their tweets are personal updates. Malicious entities, on the other hand, are trying to gain the attention of users by tweeting frequently. They also include links in their tweets and @-mention random users. The behavioral traits of these two sets of users are significantly different. Therefore, we combine PageRank with behavioral traits to derive a combined reputation value that provides a good indication of whether a user is legitimate or malicious.

Intuitively the combined reputation value reflects the dichotomous nature of Twitter as both a social network and an information dissemination site. Entities who use it for social networking have certain distinguishing behavioral traits and entities who use it to provide information to a large set of users can be judged by the quality of their true followers (PageRank). Entities who aggressively try to disseminate information and yet do not have high quality true followers do not provide useful information and are potentially malicious.

Most of the aggressive, malicious entities can be caught using the combined reputation value. However, some malicious entities in Twitter use a variety of techniques to appear legitimate which include inflating follower counts and interspersing malicious tweets with famous quotations. Many of these malicious entities are successful in having other legitimate users follow them and exhibit less aggressive behavior. We observed that despite the inflated follower counts, many of them do not have an actual SN when they enter the system. The increased follower counts are often a result of randomly following other malicious entities and getting them to reciprocate. Therefore, we can identify such sophisticated malicious entities by determining if a large set of their initial friends are malicious entities of the more aggressive kind. We correct their reputation value by decreasing it based on the number of initial friends who are aggressively malicious and the corrected reputation value is a true indication of whether they are legitimate or malicious.

In this paper we make the following contributions:

- **We combine true social network PageRank and behavioral traits to create a robust mechanism to identify malicious entities:** Our algorithm uses the dichotomous nature of Twitter to its advantage by realizing that malicious entities have poor quality information and yet try to disseminate it aggressively, thereby providing a robust identification mechanism for them. In addition, since the TSN is far smaller than the advertised SN, we are able to calculate the reputation value several orders of magnitude faster than calculating PageRank with the advertised SN. For 278,758 Twitter users, TSN PageRank takes 5 minutes while the advertised network PageRank takes 10 hours, $\approx$ 120 times slower.
- **Our algorithm identifies massively colluding malicious entities using temporal pulldown:** We are able to identify malicious entities who massively collude to appear legitimate by realizing that many of their initial set of friends are malicious entities that are more easily identified based on their behavior.
- **We validate the effectiveness of our algorithm in identifying malicious entities using real data:** We collected data for 278,758 users which included 10 known malicious entities and their social network. Our algorithm identified all the 10 malicious entities we introduced. In addition, 181 of the accounts that we identified as malicious were subsequently blocked by Twitter over a period of 31 days from March 14th to April 14th. We further identified 5334 accounts that had links to unsafe websites. To the best of our knowledge, our paper is the first attempt in automatically and efficiently identifying a broad class of malicious entities in Twitter.

In Section 2, we describe how we collected a representative sample dataset and briefly highlight insights that were gained from it. We then outline our scheme and present the concrete algorithm that results from it in Section 3. Section 4 describes our evaluation approach and discusses the key results. Related work is presented in Section 5 and the paper is concluded in Section 6.

2. DATA COLLECTION AND ANALYSIS

Category	Popular User	Sample Keywords
	NASDAQ	Dell, Coke
Science/Tech.	NASA	iPAD, Firefox API, startup
Sports	Serena Williams	winter olympics, football, basketball
Food/Health	MSNBC Health	H1N1, viral fever, dining
Entertainment	Ashton Kutcher	Shutter Island, radio, actor
News/Politics	The Economist	Chile earthquake, Joe Biden, healthcare
Education	Danah Boyd	university students, alumni, professor
Arts/Travel	Paulo Coelho	Harry Potter, Mayan pyramids, Grand Canyon
NGO/Environ.	Gates Foundation	Greenpeace, warm earth

Table 1: Categories for seed users and details behind user selection. Many of the popular users had over 1M followers, e.g., Serena Williams. The keywords represent both common and recently popular keywords used in that category, e.g., football and winter olympics.

In order to collect a representative and diverse set of users, we first picked 100 seed users from Google News categories as shown in Table 1. The number of followers for any Twitter account ranges from zero followers (e.g., newly created accounts) to over four million followers (e.g., Ashton Kutcher) and roughly follows a power law distribution. We use a similar distribution for our seed users by creating 6 buckets based on the number of followers: 0 - 250 (1), 250 - 2000 (2), 2000 - 20K (3), 20K - 100K (4), 100K - 1M (5), 1M+ (6). We ensure that the seed users per category are represented from low to high buckets in a similar power law distribution as shown in Figure 2. Each category has a mix of six

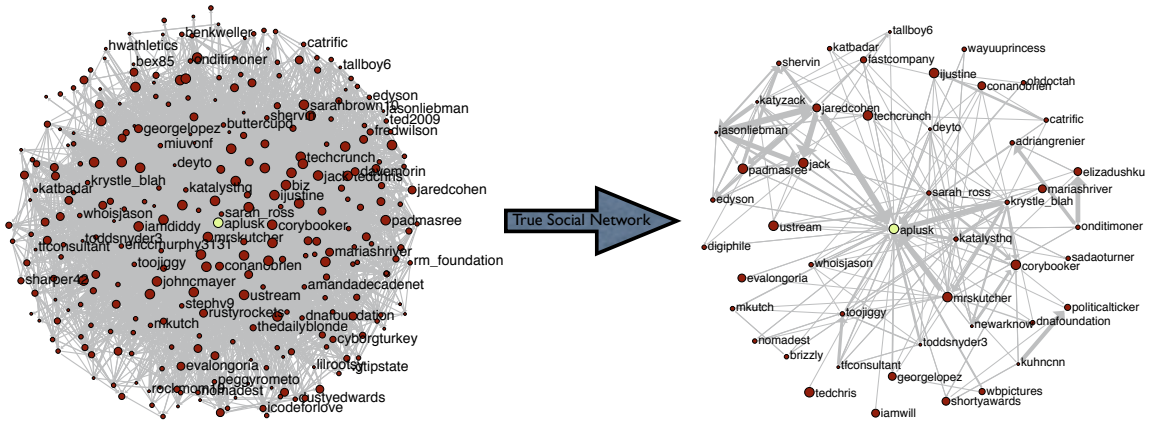

Figure 1: Out of the cluttered advertised social network of Ashton Kutcher (aplusk) emerges his true network. The node size is based on the follower count, and the edge size is based on the number of @-mentions to that particular user. The interaction between Ashton Kutcher and his wife Demi Moore (mrskutcher) is high.

bucket 1, two bucket 2, one bucket 3 and one bucket 4, 5 or 6 user. The nine users from the lower buckets 1, 2 and 3 were chosen by identifying users who were tweeting about topics related to that category. The keywords used per category are shown in Table 1. For example, in sports we picked users who tweeted about the winter olympics, football or basketball. The one user from high follower bucket (4, 5 or 6) is a very popular seed user from that category (over 20K followers) and is shown in Table 1. In sports, we picked Serena Williams who had over 1M followers. We also ensured that the tweets of each of the seed users were publicly available, recently posted and had no noticeable malicious activity.

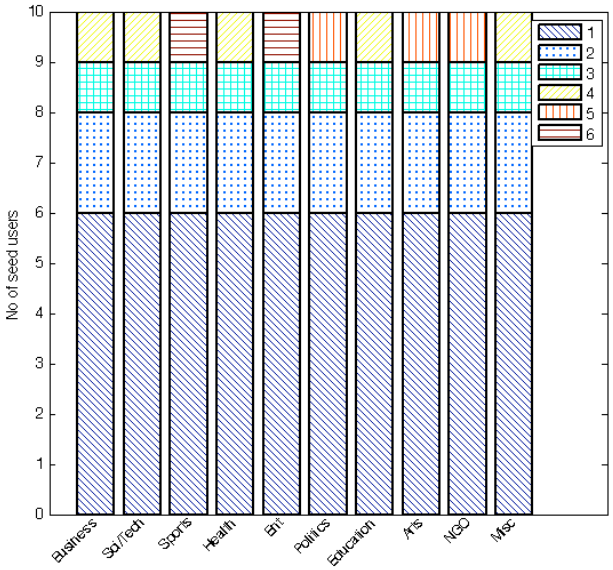

Figure 2: The seed user distribution across different categories.

For each seed user we sort the friends of the user (people that he follows) based on who he @-mentions the most. These are people the user frequently interacts with and this helps identify the true social network (TSN) of a user as opposed to what is advertised by Twitter in the form of friends and followers. For Ashton Kutcher, his advertised social network and TSN is shown in Figure 1. Though Ashton has an advertised friend list of over 300 friends, the users he interacts with form a less cluttered graph. In addition, if we look at the number of @-mentions, indicating frequency of interaction, a more closely knit group emerges that includes his wife, Demi Moore (mrskutcher), his production company Katalyst Films (katalysthq) and his animated cartoon site Blah Girls (krystle_blah). The TSN, therefore, helps in capturing the more meaningful relationships of a user.

To capture the TSN, we pick the 30 friends whom the user has @-mentioned the most and then pick 20 of their friends with the most @-mentions and similarly 10 friends for each friend of a friend. If any user has @-mentioned less than 30, 20 or 10 (based on where we are) of his friends, we pick friends at random. In addition, we only pick friends who have tweets that are publicly available. For each user, this should give his true social network within 3 hops, giving more importance to the closer social network. For each seed user and his TSN, we collect details which include name, screen name, friends list and follower list, and the last 200 tweets. In this fashion, we collected data for 252, 189 users with the Twitter Representational State Transfer (REST) API. The REST API allows 20000 requests per hour for each whitelisted account and we requested 10 such accounts from Twitter allowing us 200000 requests per hour. We included rate limiting in our scripts to ensure we made at most 150000 requests per hour. In addition, Twitter takes longer to provide friends, followers and tweets of users, due to which we collected this data over a period of two days from March 5th to 6th, 2010.

2.1 Malicious Entity Data

In addition to the seed users, we similarly collected data for 10 users who engage in various types of malicious activity as defined by the Twitter rules [8]. These predominantly include spammers who post misleading tweets or provide links to well known scams like Acai Berry supplements [1] or a mom's teeth whitening discovery [7]. We also included users who provide links to websites that host malware or indulge in phishing. For the 10 malicious accounts, we collected their TSN on March 8th, 2010 and this added 26, 569 users to our existing dataset giving a total of 278, 758 unique users.

Within our complete dataset, in addition to the 10 malicious seed accounts, we assume there exist many more malicious accounts. To verify this, over a period of 31 days from March 14th to April 14th, every night at 12 am, we ran a script to see if Twitter had suspended any account from our dataset. We found 2 of our 10 malicious seed accounts and 220 other accounts had been suspended for strange activity. These 222 accounts form our malicious account subset against which we test the effectiveness of our approach.

2.2 Representativeness of Our Dataset

To show that our dataset is representative, we plot the complementary cumulative distribution function (CCDF) of the fraction of users who have friends and followers above a certain number, and this is shown in Figure 3. This graph is plotted for all 278,758 users in our dataset. The follower line fits a power law distribution with exponent $\gamma = 2.306$ for users with number of followers lesser that 10^5. This is in accordance with analysis conducted on all of Twitter in [23] where the CCDF of the number of followers fit a power law with exponent $\gamma = 2.276$. The friend line is similar with exponent $\gamma = 2.116$. We also observe that there is a glitch in the friend line at 2000 which is the artificial friend limit imposed in Twitter [4].

2.3 Approaches for Characterizing Malicious Activity

Malicious entities succeed only if they gain the attention of other users to their substandard products or are able to direct them to their malware/phishing sites. Based on this, we can try some simple mechanisms to distinguish between legitimate and malicious accounts. However, we find that none of these mechanisms are able to make this distinction effectively. To demonstrate this, we show the results of two such methods.

We first try to use tweet frequency of users (number of tweets per day), as malicious entities who are trying to gain the attention of other users will potentially tweet far more aggressively. Figure 4 shows the tweet frequency, for (a) all users, and (b) subset that contains users who were suspended by Twitter. Since we parse only the last 200 tweets, the tweet frequency is at most 200 (the actual frequency can be higher). From Figure 4, we see that most users tweet less than 5 times a day (over 75% of users). Contrast this with only 35% of malicious accounts who tweet lesser than 5 times (Note: Figure 4 (a) is log and (b) is linear in the number of users). This shows that malicious entities are more aggressive than regular users. However, there is no single threshold which can distinguish (a) from (b), as there are some legitimate users at each tweet frequency. Surprisingly, there are a significant number of people who even tweet 200 times a day (1408 such accounts). These include 22 malicious accounts shown in Figure 4(b). We manually looked at some of the remaining 1387 accounts to see if they are all malicious. We find that there are many legitimate users who tweet 200 times a day with inspirational quotes or use Twitter as a conversational tool to constantly interact with their social network. Therefore, tweet frequency alone cannot be used to distinguish between legitimate and malicious accounts.

@-mentions are used to address a tweet to a particular user [18]. Legitimate users typically @-mention their friends who in turn interact with each other. Malicious entities

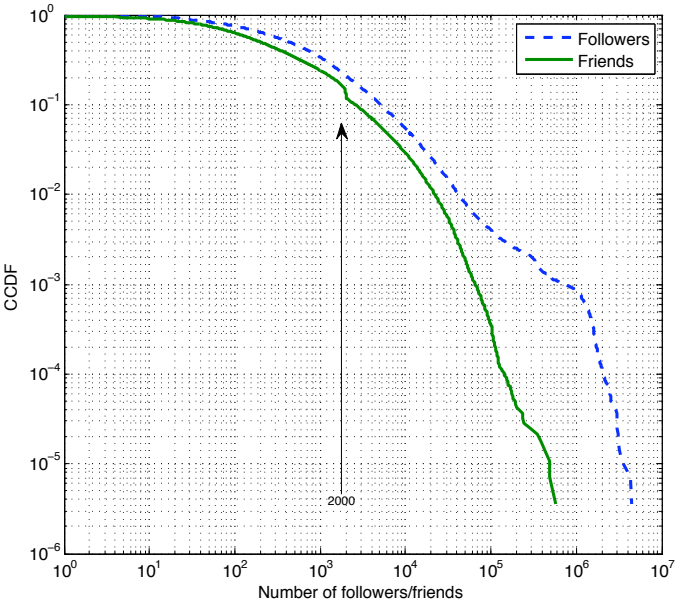

Figure 3: CCDF of the fraction that has more than a particular number of friends and followers in our Twitter sample. The graph is similar to ones plotted over the entire Twittersphere. We also notice the drop in the graph at 2000 friends due to the artificial friend constraint specified in the Twitter rules.

on the other hand try to gain attention by @ mentioning random users. Therefore, a second approach in identifying a malicious entity is seeing if a large fraction of people that it @-mentions do not interact between themselves. This can be measured by the clustering coefficient (CC) of a user. We consider our dataset to be a graph, where each vertex is a user and there exists a directed edge between user i and j if i @-mentions j. The CC of a user is defined as $\frac{\text{No. of triangles}}{\text{No. of wedges}}$. This, essentially, measures for each pair of users j and k that user i @-mentions (wedge), whether j @-mentions k or k @-mentions j (forms a triangle). Out of our sample of 278,758 Twitter users, only 211,209 users have used @-mentions in their tweets. Similarly, only 91 of the 230 malicious accounts have used @-mentions. For these users the CC is shown in Figure 5 where (a) is for all users, and (b) is for the malicious account subset. There are $\approx$ 17% of all users who have greater than .9 CC, while only $\approx$ 11% of malicious entities have such a high CC. On the other hand, both malicious entities and legitimate users have $\approx$ 4% of users with zero CC, where none of the users they @-mention ever interact. Though legitimate users have a much higher CC than malicious entities, the CC values are well spread across all values, implying malicious entities can appear as well connected as legitimate users.

3. IDENTIFYING MALICIOUS ENTITIES

As in any other system, malicious entities in Twitter are looking to draw attention of other users towards poor quality information (e.g. phony products) or to malware/phishing sites. They aggressively push this information to increase their chance of reaching some user who will be convinced to look at it. To convince the user, they also need to appear legitimate. This reveals three characteristic traits of a mali-

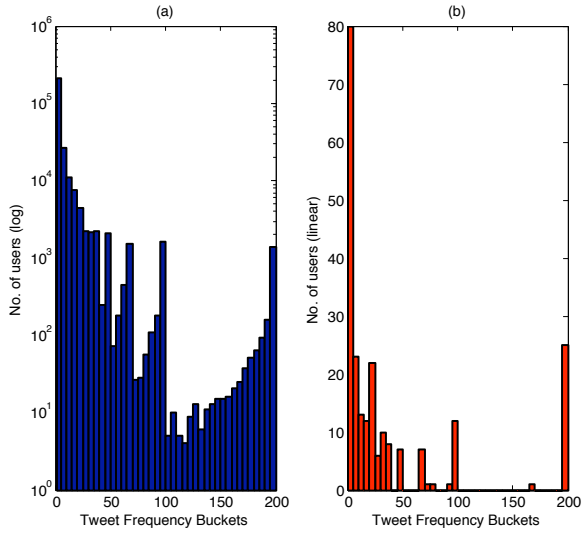

Figure 4: Tweet frequency of (a) complete user set, and (b) malicious account subset. There is no single cut-off point to distinguish legitimate from malicious accounts.

cious entity: they (1) possess substandard information, (2) make up for lack of quality by quantity, and (3) try hard to appear legitimate. The abundance of information in Twitter allows many of these malicious accounts to continue operating without being detected (information overload). However in this section, we show how we can exploit these traits to identify these accounts.

3.1 Assessing Information Quality

Twitter works both as a social network and an information dissemination site. Relationships between users are created by the process of following. When a user A follows another user B, A becomes B's follower and B becomes A's friend. B in turn can follow A to make the relationship mutual. However, this process is artificial as many users automatically follow back users who follow them including verified accounts like BarackObama. We call this the advertised social network of Twitter. Malicious entities take advantage of this, by actively following users and un-following users that do not reciprocate. Huberman et. al. [19] show that real social network engagement exists, instead, in the form of @-mentions, retweets and direct messages. Retweets (RTs) are used to repeat a tweet, and are a way to spread or popularize someone else's tweets [15]. RTs use the same notation as @-mentions, with the exception that the @ is preceded with an RT. Directed messages are similar to @-mentions but are private between the interacting users. Among the publicly available Twitter data, @-mentions mostly facilitate social networking and RTs facilitate information dissemination. We call the network gleaned from @-mentions and RTs as the true social network (TSN) of users.

PageRank has been used to judge the relative importance of a webpage, where pages that have incoming links from other highly ranked pages receive a high rank themselves (through an iterative computation). We use PageRank built on top of the TSN to rank users based on their ability to either engage other users or provide meaningful information and denote it by PR^{TSN}. Specifically, we use a weighted

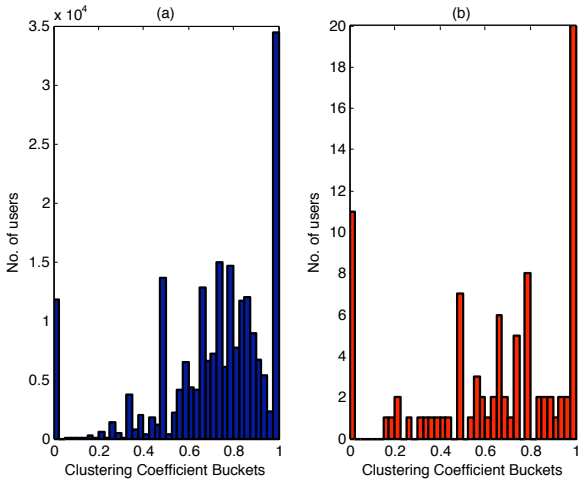

Figure 5: Clustering coefficient of (a) complete user set, and (b) malicious entities. There is no single cut-off point to distinguish legitimate users from malicious entities.

PageRank model where the link from user A to user B is weighted based on the number of times user A has either @-mentioned or RTed user B. Therefore users who are @-mentioned or RTed by well ranked users, become well ranked themselves. Formally, the transition matrix for the TSN PageRank, PR^{TSN} is of the form:

$$r_{ij} = \frac{\sum(@_{ij} + RT_{ij})}{\sum_{\forall k}(@_{ik} + RT_{ik})} \quad (1)$$

Since malicious entities have low quality information, a large number of them will end up with low PR^{TSN} values. However, not all users with low PR^{TSN} are necessarily malicious. For example, users with moderate to small sized social networks use Twitter to occasionally provide personal updates to their immediate social network. These users will not have well ranked users either @-mentioning or RTing them resulting in low PR^{TSN} values. On the other hand, malicious entities, in trying to get noticed, will produce large quantities of information and yet will not have high PR^{TSN} values. Therefore, PR^{TSN} in combination with behavioral characteristics is a good indication of whether a user is malicious.

3.2 What Constitutes Bad/Aggressive Behavior ?

In Twitter, malicious entities try to gain attention in one of the following ways: (1) tweeting frequently, (2) providing links to their products, and (3) @-mentioning random users. When users search on Twitter, the results are ordered based on time. So malicious entities tweet often to ensure that their tweets are constantly seen. Many malicious entities who are trying to link to phony products or malware will tweet frequently. Twitter policy states that a user may be considered in violation of its spam rules, *"If your updates consist mainly of links, and not personal updates"*. However, there are many legitimate entities that use Twitter primarily as a information dissemination site and tweet frequently with links in their tweets. For example, TweetMeme tweets 40 times a day on the most popular links on Twitter (digg for Twitter). The difference is that most Twitter users find

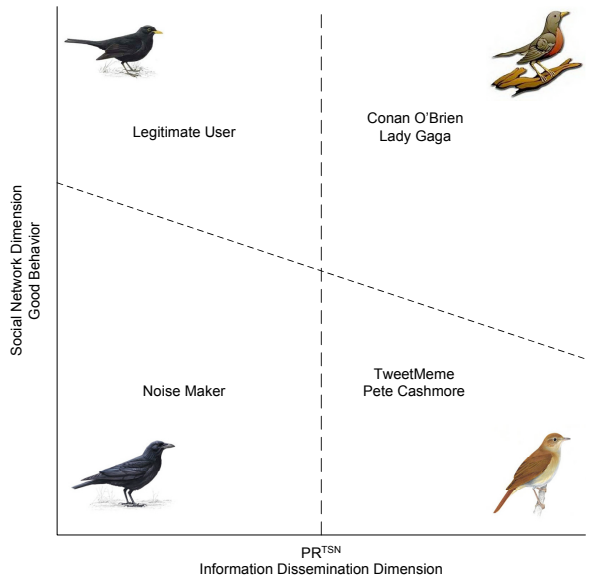

Figure 6: Users of Twitter categorized into four areas. Top left (blackbirds) are legitimate users with good behavior and low PR^{TSN}, top right (robins) are celebrities, bottom right (nightingales) are frequent event updaters (news, memes) and the bottom left(crows) are malicious entities who have low PR^{TSN} and also exhibit bad behavior.

this service informative and RT TweetMeme often, giving it a high PR^{TSN} value (rank 1).

We see that there is an interplay between PR^{TSN} and behavior, where users with low PR^{TSN} (small SN) are not necessarily bad as long as their behavior is good. Furthermore, users who appear to have poor behavior traits may not necessarily be bad as long as they have a high PR^{TSN}. However, when users aggressively try to disseminate information and yet do not have high PR^{TSN} values, they have low quality information and are potentially malicious. Intuitively, these two aspects represent the dichotomous nature of Twitter as a social networking and an information dissemination site and can be visualized as shown in Figure 6.

In Figure 6 users are categorized into one of four areas. Users with high PR^{TSN} and good behavior (top right, robin) are the celebrities who have a significant number of users interacting with them through @s or RTs. For example, Conan O'Brien, a talk show host, tweets infrequently (tweet frequency: 1.08/day), however each of his tweets are RTed many times across a wide user base. In addition, a large number of users @-mention him in their tweets resulting in a high PR^{TSN} value (7th highest). We observe that a malicious entity cannot replicate this success by getting a large number of other malicious entities to RT its tweets as PR^{TSN} inherently provides more weight to RTs from users who themselves have high PR^{TSN}. On the other hand, entities like TweetMeme which tweet 40 times a day, can afford to behave aggressively as they are also popular across a wide user base. Such entities occupy the bottom right (nightingale) area in Figure 6 along with other news entities such as Pete Cashmore's Mashable and CNN Breaking News. Legitimate users with moderate to small sized social networks have moderate to low PR^{TSN} and good behavior

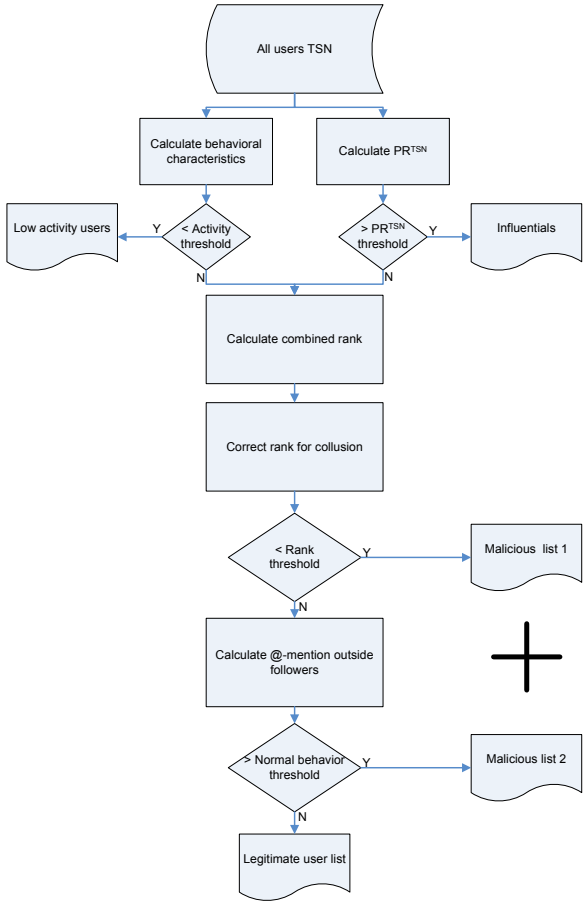

Figure 7: Combined algorithm used to identify malicious entities. It uses PR^{TSN} and behavioral characteristics to create a combined rank which is used to distinguish legitimate accounts from malicious ones

and occupy the top left (blackbird) area while the malicious accounts occupy the bottom left (crow) and are characterized by a low PR^{TSN} value and they exhibit bad behavior. Therefore, our system looks at a combination of a user's PR^{TSN} and behavioral characteristics to determine if he is legitimate or malicious.

Finally, certain malicious entities @-mention random users to get their attention. This is against Twitter's policy which states that a user can be considered a spammer, *"If you send large numbers of unsolicited @-replies in an attempt to spam a service or link"*. For example, if we make the assumption that following a user is an indication of soliciting interaction, then the metric $1 - \frac{\text{No. of @-mentions to followers}}{\text{Total no. of @-mentions}}$ shows what fraction of a user's @-mentions are unsolicited. Though following a user does not necessarily imply solicitation, this definition allows easy calculation of the metric without requiring extra data collection. We flag down a user that has a high number of outside @-mentions as a malicious entity.

3.3 Collusion Among Malicious Entities

Most malicious accounts can be identified using a combination of PR^{TSN} and behavioral characteristics. However, we notice certain malicious entities have amassed a large social network (over 5000 friends and followers), and are not as aggressive in posting links or tweeting frequently. These

entities have managed to attract some legitimate users to be part of their social network and behave less aggressively to ensure these users do not un-follow them. A user can easily be fooled into believing the legitimacy of such an account. However, we notice that many of the initial friends of such entities are actually aggressively malicious. Therefore, we can identify the more sophisticated malicious entities based on the company they start out with. The only reason they have an inflated social network is due to the more aggressive malicious accounts. On the other hand, legitimate users will rarely follow these aggressive malicious accounts. For every user, we can determine how many users in their initial social network are aggressively malicious and if this number is significantly high, identify the user as potentially malicious.

3.4 Combined Algorithm

We combine our techniques into the final algorithm shown in Figure 7. To compute PR^{TSN}, we first create the PageRank transition matrix which is a square matrix of order N, where N is the number of users. We increase the count at location (i, j) in the matrix whenever user i @-mentions or RTs user j in a tweet. After we parse all the tweets for all users, we normalize the matrix and calculate the left principle eigenvector (PageRank) of this matrix. This vector consists of reputation values, one for each user. This value is an indication of how influential a user and his tweets are. We remove users who have a high PR^{TSN} (top 5%) and categorize them as influentials [23, 26] and do not consider them in the rest of the algorithm.

Before calculating behavioral metrics, we first remove all users who have tweeted less than a certain number of tweets, t^m ($t^m = 10$ in our system) during their entire lifetime. These accounts are most often inactive or occasionally belong to a new user. For such accounts, none of the metrics we calculate are meaningful. We assume that if we need to identify a malicious account in this set, we can use other techniques like content filtering as there are a low number of tweets to be parsed. For the remaining users, we first calculate the number of tweets that have links. We use this to calculate the link ratio, $LR = \frac{\text{No. of links}}{\text{No. of tweets}}$. We also calculate the tweet frequency of the user using $TF = \frac{\text{No. of tweets}}{\text{Time between last and first tweet}}$. LR is between 0 and 1 and since we consider the last 200 tweets for a user, TF is between 0 and 200.

We calculate a combined rank based on the user's PR^{TSN} and behavioral metrics as follows:

$$CR_i = \frac{PR_i^{TSN}}{LR_i \times TF_i}$$

The denominator, $LR_i \times TF_i$ is essentially the number of tweets per day that have links. This value indicates how aggressively user i is trying to disseminate his information. PR_i^{TSN} indicates how interested other users are in propagating i's information. Users who tweet infrequently and have few links and yet get their information disseminated are users who provide naturally interesting information within the 140 character tweet. These include celebrities like Conan O'Brien whose day to day updates are interesting to a large demographic of users. These users have the highest CR values. For entities like news or meme sites, CR values are a tradeoff between their aggressiveness and how useful other users find their links (and therefore RT them). For legitimate users who only provide social updates and do not have a widespread reach, their low PR^{TSN} values are offset (explained) by their good behavior. The users with the lowest CR values are those who try to propagate information aggressively but their information is deemed worthless by most users. Malicious entities in Twitter fall in this category and we classify users below a certain CR value threshold, CR^{ls}, as malicious.

The next step of the algorithm is using the malicious entities identified above to determine if there exist collusions which contain a significant number of them and some less aggressive malicious entity. We only consider users below a certain CR value threshold, CR^{hs}, resulting in a band of users between CR^{hs} and CR^{ls} whose CR value we correct for collusions. We calculate the corrected combined rank for each user i who falls into this band as

$$CR_i' = \begin{cases} \beta \times \frac{CR_i}{N^2} & N > 0 \\ CR_i & N = 0 \end{cases}$$

where N is the number of initial friends who we identify as malicious based on CR. Essentially, when a user has malicious friends, we correct his combined rank with the number of malicious friends he has made initially.

$$\forall i | CR^{ls} \leq CR_i \leq CR^{hs} : CR_i = min(CR_i, CR_i')$$

After the pulldown all the users who have a rank below CR^{ls} are categorized as malicious and this forms the first spam list as shown in Figure 7. The first spam list contains users who have $CR_i < CR^{ls}$.

Finally, we consider the users who @-mention a large number of users outside their social network. Since @-mentions are primarily used for interaction within a user's social network, such behavior is in itself anomalous and it does not matter what their PR^{TSN} values are. Users who satisfy the following inequality are categorized as malicious:

$$\frac{\text{No. of @ mentions to followers}}{\text{Total no. of @ mentions}} \leq \gamma \times \frac{\text{Total no. of @ mentions}}{\text{Max no. of @ mentions}}$$

Since we only consider 200 tweets, we set the Max no. of @ mentions to this value. The equation ensures that users with lesser @-mentions require a greater percentage of them to be unsolicited to be caught. For example, for $\gamma = .9$, a user with 100 @-mention needs more than 55 of them to be outside his social circle (to non-followers), while a user who has only 20 @-mentions will need 18 of them to be outside to be considered malicious. As shown in Figure 7, this outputs the second list of malicious accounts and along with the first list constitutes all the users who are categorized as malicious. The rest of the users are categorized as legitimate users. In the next section we detail how we evaluate this algorithm.

4. EVALUATION

4.1 True Social Network PageRank

To compute PR^{TSN}, we first represent the transition matrix as a sorted coordinate edge list file with each line of the file of the form *(row, column, value)*, where *value* is the number of times the first user (*row*) @-mentions or RTs the second user (*column*). This file contains 3206295 edges for the 230,959 users in our system. The remaining 47799(278,758 − 230,959) users have never been @-mentioned or RTed and have never used @-mentions or RTs

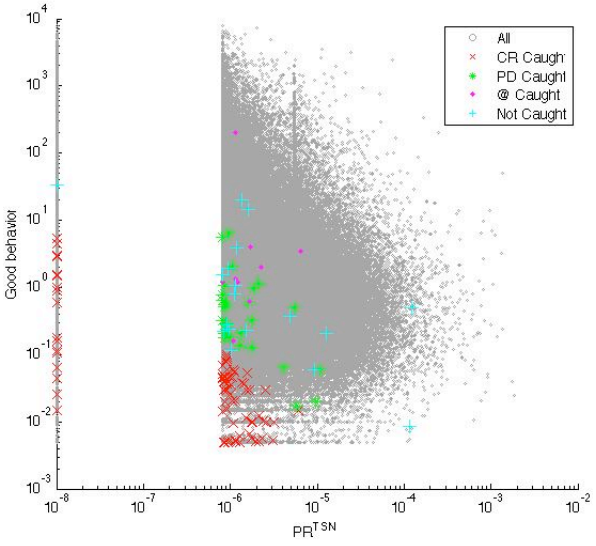

Figure 8: Users behavior against PR^{TSN}. The suspended accounts caught by the CR mechanism are shown in the bottom left and far left. Above and to the right are the accounts caught due to pulldown and @-mention mechanisms.

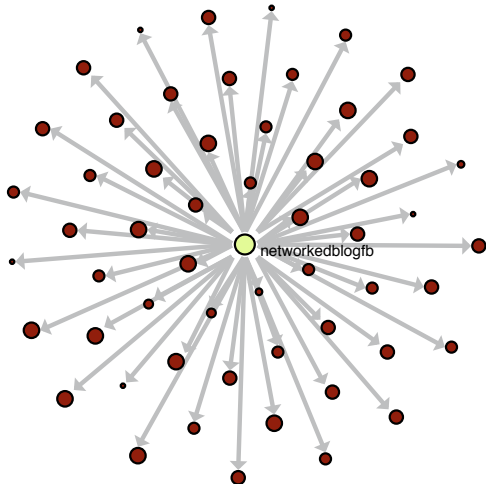

Figure 9: An account, networkedblogfb, suspended by Twitter whose initial set of friends had 56 users that the CR value identified as malicious. The size of each node is based on its CR value. Though networkedblogfb has a much higher CR value, we can identify it as malicious based on the company it starts out with.

themselves and are not considered in the PR^{TSN} calculation. Instead, we assign a PageRank value lower than the minimum PR^{TSN} value computed. We use Network Workbench (NWB) [6] which uses the iterative power method to compute the left principal eigenvector of our transition matrix (PR^{TSN}).

As the TSN matrix used is sparse, with 3 million entries ($\rho = \frac{3206295}{(230959)^2} = 6 \times 10^{-5}$), the running time is only 5 mins on a 8 core, 2-GHz Quad-Core AMD Opteron processor with 16 GB of memory, and running 2.6.24 Linux Kernel (Ubuntu 8.04.2). Contrast this with the advertised SN matrix which has 130 million entries, a 100 fold increase ($\rho = \frac{130459258}{(278742)^2} = 1.6 \times 10^{-3}$), and a running time of 10 hrs, a 120 fold time increase. Therefore, using PR^{TSN} allows a faster ranking of users and this can catch malicious entities more efficiently.

After calculating PR^{TSN} we identify the top 5% of users as influentials and do not consider them in the rest of the algorithm.

4.2 Combined Rank

The effectiveness of the combined rank can be visualized in Figure 8 where the X-axis is PR^{TSN} and the Y-axis is $\frac{1}{LR \times TF}$ (good behavior). The Figure is similar to the crow blackbird user categorization shown in Figure 6. Users on the far left were users who are not @-mentioned or RTed by any user in our system and have themselves never used an @-mention or RT. Out of the accounts suspended by Twitter, the red crosses indicate accounts that we identified as malicious using the combined rank mechanism and we can see them occupying the triangular area at the bottom left and the far bottom left. We identified the most number of malicious accounts suspended by Twitter using the CR mechanism (133 of the 230 accounts). Figure 8 also shows the malicious entities not identified by this part of the algorithm which include the green asterisks (identified by pull-down), the pink dots (identified by excessive @-mentions outside SN) and cyan pluses (not identified).

4.3 Identifying Collusions: Temporal Pulldown

Though the combined rank is successful in catching aggressive malicious entities, there are other malicious entities (green asterisks in Figure 8) who have significantly higher PR^{TSN} and they do not exhibit clearly bad behavior. A majority of these entities have attained high PR^{TSN} values by convincing legitimate users to interact with them, through their inflated friend and follower counts. However, probing into the social network history of these accounts reveals that a large number of their initial friends exhibit malicious behavior in aggressive fashion. Our pulldown mechanism exposes this and hence we are able to catch even those malicious users who have significantly higher PR^{TSN} or behavior values. These users are more sophisticated as they have better deception tactics which include better strategies at successfully amassing friends and followers when they entered the system. However, a large number of their friends and followers fall in the bottom left of Figure 8. Thus, even though these accounts may be in the far right or high above, they will get pulled down due to the company they started out with when they entered the system.

To further illustrate, we consider account *networkedblogfb* who had 17000 friends and 15913 followers before it was suspended. It posted tweets to sell traffic and followers to users which is against Twitter's spam rules [8]. It interspersed these tweets with a large number of famous inspirational quotes giving the impression that it was a legitimate account with good information and a large follower and friend base. However, our algorithm looks at the earliest set of friends made by this account and we find that 56 of its initial friends were all malicious (see Figure 9) that we previously identified using the CR mechanism. Since such entities were created for malicious activity, they do not have any real social network and therefore have to depend on a collusion to

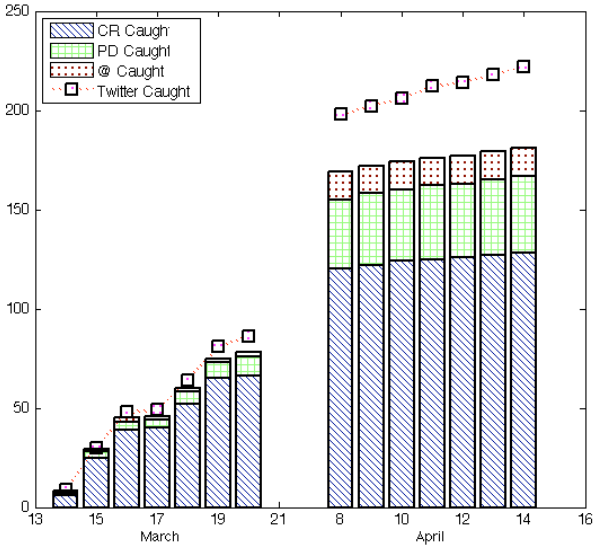

Figure 10: Number of malicious accounts in our sample that the algorithm identified and were subsequently suspended by Twitter over a period of 31 days. Also shown are all the entities in our sample that were suspended by Twitter in the same timeframe.

appear legitimate. The temporal pulldown uses this to its advantage and allows us to identify even such sophisticated malicious accounts.

4.4 Overall Effectiveness

We consider the corrected combined rank for all users which is calculated as shown in Section 3.4 and identify all users below a certain threshold as malicious. Since it is estimated that $\approx 12\%$ of Twitter users are malicious [11], we conservatively set the threshold such that 10% of users in our system are identified as potentially malicious. We found that the 10 seed malicious accounts that we considered are in the bottom 10%. After running our algorithm, we also observed the number of accounts that were suspended by Twitter over the next 31 days. The number of these accounts that were below our threshold value and the results are shown in Figure 10. Since we finished our data collection and ran our complete algorithm on March 8, and we started checking for suspended accounts only on March 14, the first data point contains all users suspended from 8th to 14th. During this time 9 accounts were suspended and we already identified 8 of them as potentially malicious. By April 14, 222 accounts were suspended out of which we identified 181.

Figure 10 also shows which part of the algorithm, the combined rank, the corrected combined rank (Malicious List 1 in Figure 7) or anomalous @-mentions (Malicious List 2 in Figure 7) identified the user. We see that the combined rank is most effective in identifying malicious accounts, followed by the temporal pulldown mechanism. This also follows from the fact that the accounts caught by the pulldown mechanism collude, are more deceptive and are fewer in number.

4.5 Users Identified as Potentially Malicious

Although 12% of Twitter accounts are estimated to be malicious [11], Twitter has so far suspended a very small fraction of them. In the bottom 10% of the users ranked by our algorithm, there are $25,272$ accounts. We observed that most of the $25,272$ users had links to a web page in their tweets. Considering 20 links at random for each user, we used Web Of Trust (WOT) [13] to classify the urls as good or bad. For shortened urls we converted them to original urls before querying WOT. WOT returns a four tuple containing ratings for 'Trustworthiness', 'Privacy', 'Vendor reliability', 'Child safety', for each url. For each of the three features other than 'Child Safety', we classified the web page as bad if the rating for the feature was less than 40%. For each web page with 'Child safety' rating less then 40%, we classified the webpage as potentially malicious. Of the $25,272$ users, 5178 had links to bad web pages and 156 had links to potentially malicious web pages in their tweets.

We used content parsing on the tweets to categorize the remaining $19,938$ users. Searching for a list of keywords and combination of keywords which include 'followers for money', 'get followers', 'twitter train'..., we could categorize 5300 users as potentially malicious. Based on the information in the 'Bio' field of the remaining $14,638$ users, we classified 1157 users as online marketers. There is a fine line that separates online marketers from malicious users. We observed that many users who identified themselves as online marketers had links to webpages that were potentially malicious, but the web pages were not identified as malicious by WOT. For the remaining $13,481$ users, more detailed techniques are needed to classify them either as malicious or as false positives.

4.6 False Positives

Since we cannot manually verify all of the bottom 10% of users($\approx$ 25K), we consider the follow friday feature in twitter, to identify the false positives among the bottom 10% users classified as potentially malicious by the algorithm. Any user can recommend another user to his followers using 'follow friday'. As the follow friday feature of Twitter allows any user to recommend other users, we observe that colluding malicious entities recommend one another. To counter this, we consider only recommendations by users not in the bottom 10%. A user recommended by a certain number of white listed users must also be good (with high probability) and hence should not be blacklisted. We found that there were only a small number of users (13) in the bottom 10% who had at least one whitelisted user recommending them through the Follow Friday mechanism.

4.7 Comparison With Other Systems

We compared our algorithm against two commercially available systems that grade Twitter users: Twitblocker built by Hashrocket Labs that identifies spammers in Twitter and Tweetgrade that was powered by Purewire research (now Barracuda networks). For the 10 seed malicious accounts that we considered, the comparison is shown in the Appendix. Our algorithm identifies all these users as potentially malicious while a large majority of them are given high grades by Twitblocker and Tweetgrade. For the first four users who are spammers, Twitblocker and Tweetgrade gives two of these users high A grades while user *Coop5993230* is given an A grade by both these systems. *Coop5993230* primarily tweets about Acai berry supplements known to be a scam [1]. His other tweets also link to products. The next three users either link to malware content or ask for Twitter account details. Again two of them are given high grades by

Twitter. In fact, *LibertyBiz* is currently suspended by Twitter and it yet receives high grades from these systems. On the other hand, we identify all these users to be potentially malicious demonstrating the effectiveness of our algorithm.

4.8 Robustness of Algorithm

It is well known that malicious users adapt to defenses that are put into a system. Thus, such users can change their behavior by becoming less aggressive or they can try to achieve a higher PR^{TSN}. The changed behavior will reduce their effectiveness which is not advantageous to them. On the other hand, higher PR^{TSN} cannot be easily attained because high reputation users will not @-mention or retweet poor quality information. Thus, we believe that the defenses designed by the combined algorithm are difficult to overcome for malicious users.

5. RELATED WORK

The recent popularity of Twitter has attracted much research. The social network of Twitter is unique as users follow both friends and other information sources and celebrities. Java et. al. [20] were the first to analyze the connections in Twitter over a sample of users and found that the in-degree and out-degree followed a power law distribution similar to many social networks. Kwak et. al. [23] conducted the same analysis over the entire Twittersphere which include 41 million users and found that the power law distribution only holds for users with less that 10^5 followers. Though we only use a sample set of Twitter users, we show a similar result, with our in degree and out degree distribution exponents closely matching those calculated over the entire Twittersphere [23], showing the representativeness of our sample.

The existence of a hidden sparse network in Twitter which reflects true user interactions is shown in [19]. They define this hidden network by creating links only between users who communicate with each other through directed messages such as @-mentions and *RT*. The conversation practices with respect to @-mentions and *RT*s in Twitter are discussed in [18, 27]. Our paper uses both @-mentions and *RT*s to compute the true social network of the user.

The fact that Twitter has characteristics of both a social network and an information dissemination site is discussed in [23, 20, 24]. Based on this Java et. al. [20] classify Twitter users into three main categories based on the size of their network: information source, friends and information seekers. Naaman et. al. [24] also show that Twitter users can be categorized into two main categories: me-formers (80%) who post tweets relating to themselves and informers (20%) who post tweets that are informational in nature. Krish et. al. [22] provide another characterization of Twitter and identify three main categories of Twitter users: broadcasters, acquaintances, and miscreants-evangelists. This paper recognizes the diversity in the intention of Twitters users and uses a multi-dimensional approach to rank users.

A significant body of research has tried to identify users who are most successful in disseminating information. Weng et. al. [26] propose an algorithm based on PageRank [25] and topical similarity to find topic wise influential users in Twitter. A PageRank inspired algorithm to compute the top influentials in Twitter is used in [9], and [23] compares identifying influentials by follower count, by advertised SN PageRank and by number of *RT*s. Our work, in contrast, focuses on identifying malicious accounts. None of the approaches used above are good for this problem as (1) the advertised PageRank calculated purely on friend and follower count can be easily manipulated by malicious entities, and (2) PageRank alone is insufficient in identifying malicious entities.

In [27], Yardi et. al. look at trending topic spam for a single topic in Twitter and identify spammers based on suggestive keywords used, the presence of a URL and the presence of numbers in spammer user names. Spammers can easily evade these techniques and we have observed many spammers in our sample who cannot be identified by any of these mechanisms. They also hypothesize that spammers follow other spammers (collusion) while legitimate users will follow other users. While we agree that spammers do collude, we see that many legitimate users including verified accounts like BarackObama (733,267 friends) also automatically follow a large number of users who follow them. To address this we use a temporal mechanism to identify a collusion.

The problem of identifying malicious entities is not limited to Twitter. Boykin et. al. [16] use social network analysis to detect email spam. Gyöngyi et. al. [17] build on PageRank to semi-automatically separate useful web-pages from spam. An algorithm similar to PageRank for computing the reputation of a node in a peer-peer network is described in [21]. This paper adapts PageRank to the true social network of users in Twitter and uses it with the behavioral characteristics and the collusion propensity of a user to identify if he is legitimate or malicious. To the best of our knowledge, ours is the first algorithm that automatically identifies a broad class of malicious accounts in Twitter.

6. CONCLUSION

In this paper we were able to identify malicious Twitter accounts based on the fact that they possess poor quality information and yet aggressively try to disseminate it. In addition, for more sophisticated malicious users who use collusion to appear legitimate, our temporal pulldown mechanism exposes their behavior. These two techniques help us automatically identify close to 82% of accounts that were eventually suspended by Twitter, demonstrating the effectiveness of our algorithm. Our algorithm can also be used to rank users across the entire Twittershpere and we will investigate ways of validating this in our future work. Though we believe our sample set is representative, collecting a larger user base will be useful in understanding the limitations of our algorithm and potential techniques that more sophisticated malicious users might use to evade detection. Finally, our algorithm provides a set of users who have a high likelihood of being malicious. Determining automatic ways in validating this would be extremely useful. Considering that Twitter is catching spammers at a slow rate our algorithm provides a good first step in identifying spammers.

7. REFERENCES

[1] Acai Berry Scams and How to Avoid them. `http://www.powersupplements.com/acai/acai-berry-scams.html`.

[2] CNN breaking news. `http://twitter.com/cnnbrk`.

[3] Computer world. `http://www.computerworld.com/s/article/9148878/Twitter_now_has_75M_users_most_asleep_at_the_mouse`.

[4] Following Rules and Best Practices. http://help.twitter.com/forums/10711/entries/68916.
[5] Greatest Vitamin. http://www.quackwatch.org/11Ind/lapre.html.
[6] Network Workbench. http://nwb.slis.indiana.edu/.
[7] Online Teeth Whitening Scams. http://www.teethwhiteningreviews.com/artman/publish/teeth-whitening-scams.php.
[8] The Twitter Rules. http://help.twitter.com/forums/26257/entries/18311.
[9] TunkRank. http://tunkrank.com/.
[10] Twitter. http://twitter.com/.
[11] Twitters Red Carpet Era Celebrities and Criminals. http://www.barracudalabs.com/wordpress/index.php/2010/03/09/twitters-red-carpet-era-celebrities-and-criminals/.
[12] Ways to Increase followers. http://twittertrain.blogspot.com/.
[13] Web of Trust. http://www.mywot.com.
[14] Woot. http://twitter.com/woot.
[15] D. Boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. *Hawaii International Conference on System Sciences*, 0:1–10, 08.
[16] P. O. Boykin and V. P. Roychowdhury. Leveraging social networks to fight spam. *Computer*, 38(4):61–68, 2005.
[17] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *VLDB '04: Proceedings of the Thirtieth international conference on Very large data bases*, pages 576–587. VLDB Endowment, 2004.
[18] C. Honeycutt and S. C. Herring. Beyond microblogging: Conversation and collaboration via twitter. *Hawaii International Conference on System Sciences*, 0:1–10, 1899.
[19] B. A. Huberman, D. M. Romero, and F. Wu. Social networks that matter: Twitter under the microscope, 2008. cite arxiv:0812.1045.
[20] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, New York, NY, USA, 2007. ACM.
[21] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 640–651, New York, NY, USA, 2003. ACM.
[22] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about twitter. In *WOSP '08: Proceedings of the first workshop on Online social networks*, pages 19–24, New York, NY, USA, 2008. ACM.
[23] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a Social Network or a News Media? In *WWW'10: Proceedings of the 19th International World Wide Web Conference*, April 2010.
[24] M. Naaman, J. Boase, and C.-H. Lai. Is it really about me?: message content in social awareness streams. In *CSCW '10: Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 189–192, New York, NY, USA, 2010. ACM.
[25] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University, 1998.
[26] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM '10: Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270, New York, NY, USA, 2010. ACM.
[27] S. Yardi, D. Romero, G. Schoenebeck, and danah boyd. Detecting spam in a twitter network. *First Monday, Volume 15, Number 1 - 4*, January 2010.

APPENDIX

A. LIST OF SEED MALICIOUS ENTITIES

Entity Name	Friend Count	Follower Count	Malicious Activity	Twitter status	Twitblocker's Grade	Purewire's Tweetgrade	Our algorithm
Coop5993230	1366	1222	Acai berry scam	Active	A+	A	Caught
MarieGerry	2300	2100	Acai berry scam	Blocked	F	none	Caught
ifatloss4idiots	2000	1200	Mom's teeth scam	Active	A+	C+	Caught
ebaydiscountz	858	297	Trending topic spam	Active	C	A	Caught
LibertyBiz	10110	10665	Misleading tweets	Blocked	A+	A-	Caught
knighta10	4481	4957	Malware Links	Active	B	F	Caught
poojadwivedi	14	42	Twitter account hijacking	Active	A+	A-	Caught
girlbellaforum	317	213	Pornography	Active	C	A-	Caught
ajalil	134	127	unsolicited @-mentions to celebrities	Active	A+	A+	Caught
hottweeters	1500	2700	unsolicited @-mentions to random users	Active	A+	A+	Caught

Table 2: List of seed malicious entities. Our algorithm is effective in identifying all of them while other Twitter user grading sites give them high grades.